

# Novoalign

## Reference Manual

Release 4.02.01, January 2020

## Table of Contents

1	<a href="#">Introduction</a>	3
1.1	<a href="#">Licensing</a>	3
1.1.1	<a href="#">Novoalign</a>	3
1.1.2	<a href="#">NovoalignMPI</a>	4
1.1.3	<a href="#">Trial Licenses</a>	4
2	<a href="#">User Guide and Examples</a>	4
3	<a href="#">Novoindex</a>	4
3.1	<a href="#">Index Size Calculations</a>	5
4	<a href="#">Novoalign</a>	6
4.1	<a href="#">Command Line Options</a>	6
4.2	<a href="#">Examples</a>	21
4.3	<a href="#">Description</a>	21
4.3.1	<a href="#">Base Qualities and Alignment Scores</a>	21
4.3.1.1	<a href="#">PRB Quality to Score Conversion</a>	21
4.3.1.2	<a href="#">Single Base Quality to Score Conversion</a>	22
4.3.1.3	<a href="#">Base Penalty Limit</a>	22
4.3.1.4	<a href="#">Base Quality to Penalty Table</a>	22
4.3.1.5	<a href="#">Match Reward</a>	23
4.3.2	<a href="#">Posterior Alignment Probabilities and Mapping Quality</a>	23
4.3.3	<a href="#">Mapping Quality and Alternative Scaffolds</a>	23
4.3.3.1	<a href="#">Novoalign / BWA MEM Differences</a>	25
4.3.4	<a href="#">Quality Trimming</a>	25
4.3.5	<a href="#">Adapter Trimming</a>	27
4.3.5.1	<a href="#">Single End Reads - miRNA</a>	27
4.3.5.2	<a href="#">Paired End Reads – Short Fragments</a>	27
4.3.6	<a href="#">Amplicon Clipping</a>	28
4.3.7	<a href="#">Read Quality</a>	30
4.3.8	<a href="#">Reads with Multiple Alignments</a>	30
4.3.9	<a href="#">Sequence file formats</a>	30
4.3.10	<a href="#">Output Formats</a>	33
4.3.10.1	<a href="#">Native Report Format</a>	33
4.3.10.2	<a href="#">Paired End Native Report Format</a>	35
4.3.10.3	<a href="#">SAM/BAM Report Format</a>	37
4.3.10.3.1	<a href="#">@SQ M5 tags</a>	39
4.4	<a href="#">Paired End Alignment Mode</a>	40
4.4.1	<a href="#">Scoring</a>	40
4.5	<a href="#">Alignment process</a>	41
4.6	<a href="#">Bisulphite Mode</a>	42
4.6.1	<a href="#">Bisulphite Report Format</a>	43
4.7	<a href="#">Quality Calibration</a>	44
4.7.1	<a href="#">Using Quality Calibration</a>	45
4.7.2	<a href="#">Quality Calibration and Novoalign Reports</a>	46



the server unless you use the -c option to reduce the number of threads.

2. Sequence read files in Gzip format can be processed allowing savings in file space. Output files can be compressed by piping into Gzip.
3. A 5' PCR adapter trimming function that is useful with some protocols such as Nimblegen Sequence Capture Arrays where a PCR adapter may have been left on the fragments.
4. BS-Seq, mode for alignment of reads from bisulphite treated DNA.
5. A base quality calibration function that calibrates base qualities based on mismatch rates from actual alignments. This improves sensitivity and specificity and is also useful for recovering alignments from poor quality runs of the Illumina Genome Analyser.
6. Handling of paired end reads where the fragment length is shorter than the read length and the reads have extended into adapter sequence. This function identifies short fragments with adapter by in-silico prepending adapter to each read of a pair and then aligning the two reads to identify short or overlapping fragments. If overlapping reads and adapter are identified the adapter is trimmed from the reads.

Note. A valid license file must be installed adjacent to the executables in order to enable multi-threading and other advanced options.

### 1.1.2 NovoalignMPI

Are only available with a license.

### 1.1.3 Trial Licenses

Trial licenses can be obtained from [www.novocraft.com](http://www.novocraft.com). Please state, organisation name and address with requests for trial licenses.

## 2 User Guide and Examples

Further documentation including examples, explanations of advanced features and performance guidelines can be found online at [www.novocraft.com](http://www.novocraft.com).

## 3 Novoindex

First step first. Novoalign requires the target reference sequences to be indexed prior to alignment. The index is saved to a file and can be reused and shared between multiple copies of the aligners. Index construction time is quite fast, a few seconds for a worm to several minutes for human genome so the index can be discarded and rebuilt as required.

Usage:

*novoindex options indexfile sequencefiles....*

Option	Description
-k 99	is the k-mer length to be used for the index. Novoindex will select appropriate values if either of these is not specified. Default value is $\log_4(N/20s)$ where N is genome size and s step size.
-s 9	is the step size for the index. Typical values are from 1 to 3, usually defaults to 1 or 2. Genomes larger than 4Gbp can be indexed using a

	stepsize > 1, the requirement is $N/s < 4G$ .
-m	lower case masking option. If included then lower case sequence is not indexed.
-b	<sup>1</sup> Creates an index based on insilico bisulphite treatment of the reference sequence. A double index based on C->T and G->A (complementary strand C>T) conversion is created. Alignments using an index created with -b option will be done in bisulphite mode.
-5	Add M5 tag to sequence headers in the index. This will also update any existing M5 tags.
-n <i>name</i>	Sets the an internal name for the reference sequence index. This is used in report headers and as the AS: field in SAM SQ record. Defaults to the <i>indexfile</i> name.
<i>indexfile</i>	is the filename for the indexed reference sequence generated by novoindex. By convention we use file suffix of <b>.nix</b> for normal indexes and <b>.nbx</b> for bisulphite mode indexes.
<i>sequencefiles</i>	a list of sequence files in fasta format to be included in the index.

Example, to generate an index file named 'celegans' for the sequence file “elegans.dna.fa”

```
novoindex celegans.nix elegans.dna.fa
```

The index includes a copy of the reference sequence compressed to 4-bits per base. The compressed format retains ambiguous nucleotide codes which will be scored appropriately by the alignment process. This feature is especially important for use with genomes that have high numbers of scattered ambiguous codes such as Maize, it's also useful for removing allelic bias, increasing specificity of alignments and improving accuracy of quality calibration.

When indexing k-mers with ambiguous nucleotide codes, index entries are created for all possible combinations of non-ambiguous codes. For instance if a k-mer contains an N, then 4 index entries will be stored with ACG&T replacing the N. To control possible explosion of index entries this process is limited to 32 entries per k-mer. Any k-mer that would create more than 32 entries is not indexed. This allows up to 5 dinucleotides in a k-mer.

### 3.1 Index Size Calculations

A normal index comprises three main tables:

1. A k-mer hash table of size  $4^{k+1}$  bytes
2. A sequence offset table of size  $4N/s$  bytes where N is the length of the sequences being index and s is the step size.
3. A compressed sequence file of size  $N/2$  bytes.

A bisulphite mode index comprises five tables, the first two being doubled up for the CT and GA

<sup>1</sup> Only available in licensed versions.

indexes.:

1. Two k-mer hash tables of size  $4 \cdot 3^k$  bytes
2. Two sequence offset tables of size  $4N/s$  bytes where N is the length of the sequences being indexed and s is the step size.
3. A compressed sequence file of size  $N/2$  bytes.

If lower case masking is specified any k-mer composed entirely of lower case codes will not be indexed. The lower case NA codes are still retained in the 4-bit/bp compressed sequence file.

## Examples

### C Elegans Genome

Genome size is 100Mbp, then using  $k=13$ ,  $s=1$  the index size is

$$\begin{aligned} &= 250\text{Mb} + 400\text{Mb} + 50\text{Mb} \\ &= 700\text{Mbytes} \end{aligned}$$

With  $k=13$  and  $s=3$  the size would be  
 $= 250\text{Mb} + 133\text{Mb} + 50\text{Mb}$   
 $= 433\text{Mbyte}.$

### Homo Sapiens Genome

For searching the full human genome on an 8Gbyte RAM server the recommended settings are  $k=14$ ,  $s=3$ . This gives a theoretical index size of:

$$\begin{aligned} &= 1\text{Gb} + 4\text{Gb} + 1.5\text{Gb} \\ &= 6.5\text{Gbytes} \end{aligned}$$

In practice the size is 6.0Gbytes due to N regions which are not indexed.

For searching the full human genome on an 16Gbyte RAM server the recommended settings are  $k=14$   $s=2$  or  $k=13$ ,  $s=2$ . The theoretical index size for  $k=15$ ,  $s=2$  is:

$$\begin{aligned} &= 4\text{Gb} + 8\text{Gb} + 1.5\text{Gb} \\ &= 13.5\text{Gbytes} \end{aligned}$$

Novoindex is multi-threaded and will use all available CPUs. Typical index build time for Human Genome index ( $k=14$ ,  $s=3$ ) on a dual core AMD Athlon CPU is approximately 3 minutes.

## 4 Novoalign

Aligns sequencing reads against an indexed set of reference sequences. Novoalign uses an iterative search algorithm to find the best alignment and any other alignments with similar score.

### 4.1 Command Line Options

Usage:

novoalign options

Option	Description
<b>Reference Genome Options:</b>	
-d <i>dbname</i>	Full pathname of indexed reference sequence from novoindex
--mmapoff	Disables memory mapping of the index. With this option the index is loaded into the local memory of the process. See NovoalignMPI User Guide for notes on how and when to use this option.
--hugePage	Attempt to allocate RAM for index using Huge Pages. The Linux OS should be configured with enough huge pages to fit the index. Using 1G Huge pages can improve performance >5%
	Note. --mmapoff will use anonymous huge pages if configured.
--lockidx	In memory mapped mode this will lock the index in RAM. This may improve performance in some situations. Valid on Linux servers supporting LOCK option on Memory Mapped files.
--alt	Enables alternative scaffolds mode. This changes processing for MAPQ so that surrogate mappings do not affect the MAPQ and also adds a new tag ZA:i: with a MAPQ that includes all mappings in its calculation.
<b>Sequencing Read related Options:</b>	
-f <i>seqfile1</i> [ <i>seqfile2</i> ]	Files containing the read sequences to be aligned. File formats allowed include Solexa PRB, Sanger FASTQ, FASTA, Solexa FASTQ, Illumina FASTQ, Illumina qseq_txt, and unaligned BAM files. If two files are specified then they are treated as paired end reads.
--hdrhd [ <i>9 off</i> ]	Controls checking of identity between headers in paired end reads. Sets the Hamming Distance or disables the check. Default is a Hamming Distance of not more than 1. Processing will stop with appropriate error messages if Hamming Distance exceeds the limit. This test is useful for detecting problems with ordering or missing reads in paired end fastq files.
--interleaved	When used with a single fastq input file treat file as interleaved paired end reads.
-F <i>format</i> [ <i>tags</i> ]	Specifies the <i>format</i> of the read file. Normally Novoalign can detect the format of read files and this option is not required. However starting with Illumina pipeline version 1.3 the scale for quality values has been changed. If you are using the new format Illumina *_sequence.txt files you need to add the option '-F ILMFQ' to ensure correct interpretation of quality values. Other values for the -F option are:
FA	Fasta format read files with no qualities.
SLXFQ	Fastq format with Solexa style quality values. $10\log_{10}(P/(1-P)) + '@'$
STDFQ	Fastq format with Sanger coding of quality values. -

Option	Description
	$10\log_{10}(\text{Perr}) + '!''$
ILMFQ	Fastq with Illumina coding of quality values. - $10\log_{10}(\text{Perr}) + '@'$
ILM1.8	Illumina Casava V1.8 fastq files with Sanger coding of quality values. - $10\log_{10}(\text{Perr}) + '!''$ .
BAMSE	The input file is a BAM file and all reads will be aligned in single end mode.
BAMPE	The input file is an BAM file of paired end reads in read name order. Reads will be aligned in paired end mode. Any single end reads will be skipped.
BAM	The input file is a BAM file and reads will be aligned in single/paired mode depending on flag attributes.
	<p>Notes.</p> <ol style="list-style-type: none"> <li>1. BAM files should be sorted such that the two reads of a pair are adjacent.</li> <li>2. The -F BAM* options also allow a list of SAM tags to be specified. These tags will be copied from the input BAM to the output alignments. e.g. <b>-F BAM RX,QX</b> <b>The RX tag should be on both reads of the pair for proper function of Novosort MarkDuplicates.</b></li> <li>3. BAM files should have at most one @RG record.</li> </ol>
TSV	A single line tab separated read file format.

### Notes

1. For various fastq format files, even if the -F option is used  
Novoalign will still check the actual quality values and verify  
they are consistent with the -F setting.
2. If named pipes are used for the read sequence files then the -F  
option is required.
3. FASTQ format reads with SAM output format allow copying  
SAM tags from FASTQ header comment to the SAM record.  
See the -C option.

The following three options apply to ILM1.8 format files and specifies how reads flagged as low  
quality by Illumina base caller will be processed.

- ILQ\_SKIP                      Flagged reads are not processed and not written to output reports. This  
is the default action.
- ILQ\_USE                      Flag is ignored and reads are treated as per any other read.  
Note. As these reads might be from polyclonal clusters we suggest



Option	Description
	using together with the -p option.
--ILQ_QC	Reads are written to output report with QC flags set. No attempt is made to align the read.
-# 99[K M]	Sets a limit on the number of reads or pairs to process from the input files. e.g. <b>-#10K</b> will only align the first 10,000 reads.
-# 99.9%	Specifies a percentage of reads to process. e.g. <b>-#1%</b> will process every 100 <sup>th</sup> read. This can be used with an absolute limit on the number of reads as in <b>-# 0.1% -#2K</b> will process every 1000 <sup>th</sup> read until 2000 reads have been processed.
-# x:n	Skip x reads and then map every n <sup>th</sup> read where x & n are whole numbers. Examples... -# 1:10               Skip 1 read and then map every 10 <sup>th</sup> read. -# 1000:1 -# 5       Skip 1000 reads and then map the next 5 reads.
<b>Note.</b> This and -# 99% formats are mutually exclusive and if both are given on the command line the latter will take affect. The -# 99% option is redundant and will be removed in a future release.	

### ***Alignment Scoring Options:***

-t A,B	Sets the alignment score threshold as a function of read length. threshold = (L - A) * B Where: L is read length (sum of pairs) A could be set to log4(Reference genome length). B can be fractional. Default is <b>-t 0,3</b>
-t 99	Sets absolute threshold or highest alignment score acceptable for the best alignment.
-g 99	Sets the gap opening penalty. Default 40
-x 99	Sets the gap extend penalty. Default 2
--matchreward 9	Sets a match reward. Default 4

### ***Bi-sulphite Alignment Options:***

-u 99	Sets a penalty for unconverted cytosines at CHG and CHH positions as these are less likely to be methylated than CGH sites, thus biasing alignment in favour of methylated CG sites. Default is 8. Suggested values are 12 for vertebrate and 8 for plants on 50bp reads. Using this option can reduce runtime and is only effective in -b4 mode.
-b mode	Sets Bi-sulphite alignment mode. Values for mode are: 4 - Aligns in 4 possible combinations of direction (forward & reverse complement) and index (CA & GT). (Default) 2 - Aligns reads in forward direction using CT index and in reverse

Option	Description
--------	-------------

complement using the GA index. This option is appropriate if using standard Illumina Bi-seq protocol as it preserves strand of the fragments.

### Quality Control and Read Filtering Options:

-l 99 Sets the minimum number of good quality bases for a read. Default is set to  $\log_4(Ng) + 5$  where Ng is the length of the indexed reference sequences. This test is based on information content of the read using Shannons Entropy,  $H(X) = - \sum_{i=1}^n p(x_i) \log_b p(x_i)$ .

Base Quality	Counts as .. bp
40	1
30	1
20	0.95
10	0.7
5	0.3
2	0

For good performance -l should be set around half the read length. Setting -l below (or even near)  $\log_4(N)$  where N is reference genome size will likely cause severe performance problems.

-h 99 [99] Sets a threshold for the homopolymer and optionally the dinucleotide repeat filters. All reads are checked to see if they are homopolymers (or dinucleotide repeats) and if so they are not aligned. Base qualities are used in calculating a homopolymer score. If the score is less than the threshold then the read is deemed to be a homopolymer. Default value is 20 and 120 for Bi-Seq alignments.

Setting a negative values disables homopolymer filtering, -h -1 -1 will disable both filters.

The second threshold is used for filtering dinucleotide repeats. This can useful for improving performance when aligning against genomes with high dinucleotide repeats.

For paired end reads both reads would have to be homopolymers or dinucleotide repeats for alignment of the reads to be skipped.

Reads that are over threshold are reported with a status of 'QC'

--hlimit <F> Alternative option for handling homopolymeric reads. This limits time spent trying to align reads which are primarily homopolymeric. In many cases these reads are artefacts from the sequencing process and they can take a long time to align as they seed to many locations in the genome and, when they do align, they usually have very low alignment quality.

This option limits the alignment threshold so that the maximum number of mismatches allowed is a function of the number of mismatches required for the read to align to a perfect homopolymer.

Where:

F is typically in range 5-15 and limits alignment threshold to  $F * N_h$

$N_h$  is number of mismatches required to align to a perfect

Option	Description
	homopolymer. As mismatches typically score 30 a value of 10 would allow 1/3 the number of mismatches as there are bases differing from a homopolymer. Default 8.
-H [limit [margin]]	Hard clips 3' bases with average quality <= limit from reads before aligning them. This uses a modified Mott algorithm similar to that used in BWA. Starting at 3' most base we calculate the base (quality – limit), keeping a running sum of this value. If the running sum exceeds the margin, or we reach the 5' end of the read, then bases from the minimum value to the 3' end of the read are trimmed. Hard clipping is applied before the polyclonal filter. Any N's in read are treated as quality 2. Specifying -H alone sets the quality limit at 2. The margin value defaults to 120. Default 8
--trim3HP <i>baselist</i>	Hard clip 3' homopolymers regardless of base quality. Min length 15bp and 88% pure. Useful for reads that degrade to high base quality homopolymer sequences as seen in some 2-dye runs. Applied after -H if used. Try --trim3hp AG to trim 3' either A or G homopolymer sequences from 2-colour reads. Use --trim3hp R to trim 3' mixed AG sequences.
-p 99,99 [0.9,99]	Sets thresholds for polyclonal filter. This filter is designed to remove reads that may come from polyclonal clusters or beads. Please refer to paper: <i>Filtering error from SOLiD Output, Ariella Sasson and Todd P. Michael</i> . The first pair of values (n,t) sets the number of bases and threshold for the first 20 base pairs of each read. If there are n or more bases with phred quality below t then the read is flagged as polyclonal and will not be aligned. The alignment status is 'QC'. The second pair applies to the entire read rather than just the first 20bp and is specified as fraction of bases in the read below the given quality. Setting -p -1 disables the filter. Default for is <b>off</b> .
--Q2Off	For Novoalign disables treating Q=2 bases as Illumina "The Read Segment Quality Control Indicator". Setting Q2 off will treat Q=2 bases as normal bases with a quality of 2. When off Q=2 bases are included in quality calibration and may be recalibrated to higher qualities.

### Read Preprocessing Options:

-a [ <i>adapter1</i> ] [ <i>adapter2</i> ]	<sup>2</sup> Trims 3' adapter sequence from read prior to alignment. Default adapter sequence is 'Gex Adapter 2' ,
---	--

<sup>2</sup> Adapter trimming of paired end reads is only available in Licensed versions of Novoalign. Unlicensed versions can trim adapter from single end reads for miRNA projects.

Option	Description								
	<p>"TCGTATGCCGTCTTCTGCTTG".</p> <p>e.g.</p> <pre>novoalign -a TCGTATGCCGTCTTCTGCTTG</pre> <p>This is usually used when sequencing small RNA.</p> <p>With paired end reads it can be used to trim adapter off fragments that are shorter than the read length. In this case you can specify two adapter sequences, the first for read 1 of each pair and the second for read 2. If only one is given it is used for both reads of the pair.</p> <p>Default adapter sequences for paired end reads are:</p> <p>Read1: AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG</p> <p>Read2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGA</p> <p>For Illumina mate pair reads, when both short and long fragment lengths have been entered with the -i option, the two reads from a short fragment will be trimmed to remove the adapter and the overlap. This allows proper identification of reads that overlap the circularisation junction.</p>								
-n 99	Truncates reads to the specified length before alignment. Useful for truncating reads when 3' quality is really bad..								
-s [9]	<p>Turns on read trimming for single end reads only. Reads that fail to align will be progressively shortened by specified amount (defaults to 2) until they either align or length reduces to less than the length set by the -l option, in which case the shortened read fails quality control checks. This option only applies to single end reads. Use at your own discretion.</p> <p>e.g.</p> <p>To trim reads in steps of 2 bases...                      novoalign -s</p> <p>To trim reads in steps of 5 bases...                      novoalign -s5</p>								
-5 r1[,l1] [r2[,l2]]	<p><sup>3</sup> Trims 5' primer sequences from reads before aligning. Default is not to trim 5' sequences.</p> <p>This option is useful where sample preparation protocol involved an additional PCR step with non-Solexa primers that may still be present on the 5' ends of reads.</p> <p>It can also extract Molecular Barcodes from 5' region of reads and add to SAM alignments as RX &amp; QX tags,</p> <p>This option is similar to the -a except that it acts on the 5' end of reads. It will trim partial primer sequences.</p> <table border="1"> <tr> <td>r1</td><td>primer sequence for first read of pair.</td></tr> <tr> <td>l1</td><td>minimum bp to trim, defaults to zero.</td></tr> <tr> <td>r2</td><td>primer sequence for second read of pair.</td></tr> <tr> <td>l2</td><td>minimum bp to trim from read2.</td></tr> </table> <p>Any X's in the primer sequence are treated as N's for comparison and</p>	r1	primer sequence for first read of pair.	l1	minimum bp to trim, defaults to zero.	r2	primer sequence for second read of pair.	l2	minimum bp to trim from read2.
r1	primer sequence for first read of pair.								
l1	minimum bp to trim, defaults to zero.								
r2	primer sequence for second read of pair.								
l2	minimum bp to trim from read2.								

<sup>3</sup> 5' PCR primer trimming is only available in licensed copies of Novoalign

Option	Description
--------	-------------

the matching read bases are extracted as part of Molecular Barcode.  
5' trimming is done after any 3' trimming.

#### Examples

-5 ACTACTG	Trims the primer sequence from 5' ends of read 1. Partial primer matches will be trimmed.
-5 ACTACTG,5	Trims read1 a minimum of 5bp or more if read matches the given primer sequence.
-5 ,10	Trims 10bp from 5' end of R1
-5 ,10 ,0 or -5 NNNNNNNNNN ,0	Trims 10bp from 5' end of R1.
-5 ,10 ,6	Trims 10bp from R1 and 6 bp from R2
-5 ,0 RRRRRRRRR,5	Trims 5' low complexity AG sequences up to 9bp from read2 with minimum trim of 5bp. (e.g. SWIFT Biosciences ACCEL-NGS 1S low complexity tails). We suggest using 20 R's
-5 XXXXXXXXXXX ,0	Extracts 10bp MBC from 5' of read1.
-5 ,0 XXXXXXXATTGGAGTCCT	Hardclips a 18bp MBC/Adapter pair from 5' of read 2, writing the 7bp MBC portion to the RX tag of the SAM alignment. If adapter portion does not map to read then no bases are clipped.
-5 ,0 XXXXXXXXATTGGAGTCCT,18	As above but always clips 18bp from the read.

#### Reporting Options:

-o [format | option]

Specifies output report format and options.

-o [Native | Extended]

Specifies the report format. **Native**, **SAM**, **BAM**, or **Extended**.  
Default is SAM. BAM format allows a compression level to be set.  
e.g.

-o SAM [read-group]

novobalign -o SAM

-o BAM [0-9] [read-group]

or ,

novobalign -o BAM 6 "@RG\tID:..."

When SAM or BAM format is specified a read group record (@RG) can follow the -o option. Note that the @RG record should be tab delimited and in **bash** shell you can do this using '\$'\t...' syntax. e.g.

novobalign -oSAM '\$'\tRG\tID:readgroup\tSM:sample\tPU:platform-unit\tLB:library'-d ...

Novobalign will also convert any '\t' in the option to tabs so you can also use :-

Option	Description
	<p><code>novoalign -oSAM "@RG\tID:readgroup\tSM:sampleId" -d ...</code></p> <p>ID &amp; SM fields are required by Novoalign. Note. GATK MarkDuplicates requires LB (library) to be set.</p> <p>The read group ID will be included as a tag on the alignment records as per SAM specifications. PU &amp; LB values, if present, can also be added to SAM lines as tags. See the <code>--tags</code> option.</p>
<code>-o Sync</code>	In multi-threaded mode ensures that the output report is synchronous with the read file. This may increase memory usage.
<code>--softclip 99[,99]</code>	<p>Turns on soft clipping and sets a reward for alignments extending to the start or end of a read. Typical value of 40. The first value is for 5' of read and second for 3' of read. Default 40,25.</p> <p>If only one value is given it applies to both 5' &amp; 3' ends of the read.</p> <p>The reward is only used in the soft clipping routine and is not added to the reported alignment score.</p>
<code>-o SoftClip</code>	<p>With this option alignments in SAM format will be soft clipped back to the best local alignment. On by default from V2.06.10.</p> <p>This option helps reduce SNP and micro indel noise from the ends of alignments and improves SNP specificity.</p> <p>The option can also be used with Native format to limit SNP calls to those within the best local alignment.</p> <p>Equivalent to <code>--softclip 0</code></p>
<code>-o FullNW</code>	<p>Turns off soft clipping so all bases in the read are (other than adapter trimming) reported as matches or indels. This may report inserts at the ends of reads that align across the ends of reference sequences or across structural breaks in the genome.</p> <p>Equivalent to <code>--softclip 9999,9999</code></p>
<code>--3Prime</code>	<p>Reports 3' mapping location of read. In SAM format this is tag <code>Z3:i:</code> and in Native format is an extra column immediately after the 5' mapping location.</p> <p>This option is obsolete and will be removed in a future release. Please see <code>--tags</code> option for alternative mechanism to enable the Z3 tag.</p>
<code>-R 99</code>	<p>Specifies a score difference between first two alignments for reporting repeats. If the difference is less than this then the read is treated as aligning to a repeat and '-r method' applies.</p> <p>Default is 5.</p> <p>When used with <code>-r Exhaustive</code> it increases the score range for reported alignments.</p>
<code>-r method [limit]</code>	<p>Sets the rules for handling of reads with multiple alignment locations. Values are:-</p> <p>None          No alignments will be reported. The read will be</p>

Option	Description
	reported as a status R with no alignment locations. Default except for small RNA Mode.
Random	A single alignment location is randomly chosen from amongst all the alignment results.
All	All alignment locations are reported. The 'All' method can optionally specify a limit for the number of lines reported. e.g. '-r A 10' will report at most 10 randomly selected alignments. Only alignments with score less than best alignment plus the -R setting are reported. -R defaults to 5. default if -m option is used.
Exhaustive	Reports all alignments with a P(R Ai) score less than or equal to the threshold plus the -R setting. The 'Exhaustive' method requires that a limit for the number of lines reported. e.g. '-r E 1000' will report at most 1000 randomly selected alignments per read. This is to avoid situations where high copy number repeats result in reporting millions of alignments for a read.
--2NoSQ	For secondary alignments set SEQ and QUAL to '*'
-e 999	Sets a limit on number of alignments recorded for a read during the iterative search process. The limit applies to the number of alignments with score equal to the best alignment. When limit is reached no further alignments are recorded and the search for this read is stopped. Default is 1000 in default report mode, in other report modes the default is no limit. This limit is designed to reduce CPU utilisation for reads that align to high copy number repeats and that would be reported with an 'R' status.
-q 9	Sets number of decimal places for quality score. Default zero. Example: -q2 will print quality scores with 2 decimal places.
--rNMOri	If a read is unmapped report read sequence and original qualities before any hard clipping or quality calibration.
--nonC	Sets Novoalign(MPI) to non-concordant mode. By default Novoalign reruns using same input files, options and versions should produce identical results. When set to non-concordant mode results may differ slightly between runs due to threading and application of fragment length penalties and quality calibration. In Concordant mode there are some pauses in the alignment process while threads synchronise fragment length and base quality data.
--amplicons <i>amplicons.bed [delta]</i> <i>[output.bed]</i>	Enables soft clipping of 3' & 5' ends of reads where they align to the primer sequence of an amplicon. Soft clipping will be from the start of the mapping to the end of the primer.

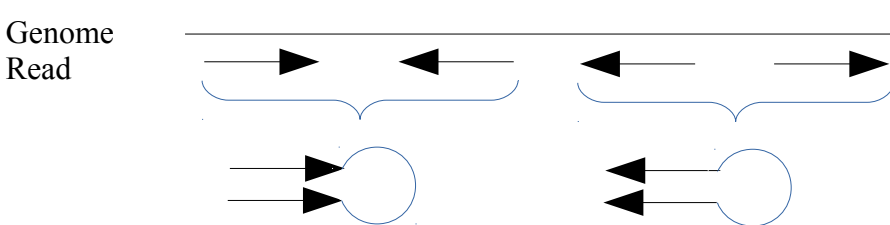


Option	Description
	<p>The optional delta field allows a fragment that maps delta bp outside the amplicon coordinates to still be detected as an amplicon match and have the primer soft clipped. Default 1, range 0-9.</p> <p>If optional output.bed field is set Novoalign writes the amplicon bed file with the score column set to the number of read pairs that matched the amplicon.</p> <div> <p>Note. An invalid value for delta will likely get treated as an output.bed filename.</p> </div>
--tags tag[-]...	<p>List of SAM tags to be enabled or disabled. Please refer to description of SAM tags.</p> <p>Example, to enable the Z3 tag and disable the MD tag...</p> <pre>--tags Z3 MD-</pre>
-C	<p>Append FASTA/Q comment to SAM output. This option can be used to transfer read meta information (e.g. a molecular barcode tag) to the SAM output. Note that the FASTA/Q comment (the string after a space in the header line) must conform the SAM specification for tags (e.g. BC:Z:CGTAC). If there are multiple tags they should be tab delimited. Malformed comments may lead to incorrect SAM output. If only one read of a pair has tags then they will be copied to SAM records of both reads in the pair.</p>
--addM5 [on off]	<p>When On add an M5 checksum tag to and @SQ records missing the tag. Default On.</p>
<b>Paired End Options:</b>	
-i [MP PE ++ +- -+] 99[ -,]99	Sets fragment orientation and approximate fragment length for proper pairs.
-i MP 99[ -,]99	MP Sets for Illumina mate pair orientation (-+)
99[ -,]99	PE Sets paired end orientation, +-.
	+ - Sets orientation where two reads of a pair are on opposite strands and facing each other. Equivalent to setting PE.
	- + Sets orientation where two reads of a pair are on opposite strands and facing away from each other. This is normal mode for Illumina mate pairs.
	++ Sets orientation where two reads of a pair are on same strand as in 454 mate pairs.
	<pre>&lt;----F3----&lt;----R3----</pre> <pre>or ----R3----&gt;----F3----&gt;</pre>
	<p>Expected fragment lengths sizes can be set as a mean and standard deviation or as a range of lengths using '-' as delimiter.</p> <p>Examples:</p>



Option	Description
-i 250 50	Defaults to paired end Illumina with 250bp insert and 50bp standard deviation
-i PE 250,50	Uses paired end orientation with 250bp insert and 50bp standard deviation
-i MP 2000,200	Uses mate pair orientation with 2000bp insert and 200bp standard deviation
-i +- 50-300	Sets +- (paired end) orientation with proper pair fragments ranging in length from 50 to 300bp. Fragment length penalties are not applied. This format should be used for amplicons where application of a fragment length penalty is not appropriate.
<p>When a range of fragment lengths is specified Novoalign will not apply fragment length penalties and this may impact ability to resolve alignments near tandem and other local repeats.</p> <p>The second form, <b>-i MP 99,99 99,99</b> , allows both a long insert length and a short insert length to be set for Illumina mate-pair reads. When this format is used Novoalign will map proper fragments of either type. It will also handle the case where the circularisation junction is within one of the reads, reporting the alignment to the longer portion of the read. In this mode the longest fragment length is 65000bp.</p> <p>Example:</p> <p>-i MP 2500,600 250,50      Specifies mixed mate pair and paired end reads.</p> <p>Proper setting of orientation is important. If in doubt about mean fragment length and standard deviation err on the high side.</p> <p><b>Default for Novoalign</b> is paired end reads with mean length of 250bp and standard deviation of 50bp.</p> <p>Changing between Paired end and Mate pair mode changes the expected orientation of the alignments in a proper pair. For paired end reads the alignments are on opposite strands and face each other</p> <pre>           &gt;-----&lt;           &lt;-----&gt; </pre> <p>For Illumina mate pairs the alignments face outwards</p> <pre>           &lt;-----&gt;           &gt;-----&lt; </pre>	
--pechimera [on off]	Enables use of supplementary alignments when one read of a pair is chimeric. Default On
-v 99	Sets the structural variation penalty for chimeric fragments. This form uses a single penalty for all pairs that do not fit the fragment length distribution. Default penalty is 150.
-v 99 99	Sets the structural variation penalties for chimeric fragments. In this

Option	Description
-v 99 99 99 regex	<p>form the first penalty applies to chimera where the alignments for the two reads of a pair lie in the same reference sequence and with orientation as per -i option.</p> <p>The second penalty applies to chimera that cross reference sequences or where orientation does not agree with the -i option.</p>
	<p>Sets the structural variation penalties for chimeric fragments. The three penalties are for:</p> <ol style="list-style-type: none"> <li>1. Penalty for SVs within a group of sequences as defined by the regular expression and orientation as per -i option.</li> <li>2. Penalty for SVs within a single sequence and orientation as per -i option.</li> <li>3. Penalty for SVs across different sequences and groups or where alignment orientation does not agree with the -i option.</li> </ol> <p>regex defines a regular expression applied to headers of indexed sequences. The regular expression should define one field that selects a group name field from the sequence header.</p> <p>This feature is often used for mRNA alignments when the reference sequence includes exon/exon junction records, to group together the exon sequences and the junction sequences by Gene id</p>

Option	Description
<b>miRNA mode:</b>	
-m [99]	<p>Sets miRNA mode. In this mode each read is given an additional score (SAM ZH tag) based on the Needleman-Wunsch alignment of the read to the opposite strand. Precursor miRNA which form hairpin structures should get a better score for the adjacent opposing strand alignment.</p>  <p><i>Precursor miRNA forms a hairpin structure which means that there should be adjacent forward and reverse complement alignments to the miRNA. Novoalign reports an additional score for the best nearby alignment on the opposite strand to the primary alignment.</i></p> <p>The optional parameter [99] controls the length of the sequence region scanned for the reverse complement alignment and is the maximum distance (gap) between the two alignments of the hairpin structure. Default is 100bp. (in earlier versions of Novoalign this was fixed at 50bp)</p> <p>In miRNA mode the repeat reporting is defaulted to 'All'. The miRNA mode does not turn on adapter filtering. This allows use with reads that have already had the adapter trimmed from them.</p> <p>This also changes default gap open to zero, gap extend to 30 and match reward to 3 with soft clipping turned off.</p>
<b>Multi threading<sup>4</sup>:</b>	
-c 99	Sets the number of threads to be used. On licensed versions it defaults to the number of CPUs as reported by sysinfo(). On free version the option is disabled.
--mCPU 99	For MPI version specifies number of threads to reserve for master process if a slave is on the same node as the master. Only applied if -c option is not used in which case any slave on same node as master gets (#Cores – mCPU) threads.
<b>Quality Calibration<sup>5</sup>:</b>	
-k [infile]	<p>Enables quality calibration. The quality calibration data (mismatch counts) are either read from the named file or accumulated from actual alignments. Default is no calibration.</p> <p>Note. Quality calibration does not work with reads in prb format.</p>
-K [file]	Accumulates mismatch counts for quality calibration by position in the read and called base quality. Mismatch counts are written to the named file after all reads are processed. When used with -k option the

4 Licensed versions only.

5 Requires a license

Option	Description
	mismatch counts include any read from the input quality calibration file.
--rOQ	If quality calibration is on then write original base qualities as SAM OQ:Z: This option is obsolete and will be removed in a future release. Please see --tags option for alternative mechanism to enable the OQ tag.
--rNMOri	If a read is unmapped report read sequence and original qualities before any hard clipping or quality calibration.
--qbin [bins]	Enables binning of quality values in alignment report. This helps reduce BAM size by limiting the values for base qualities and improving compression. This is useful with quality calibration (-k) as it increases the number of possible values and hence the BAM file size. It can also be useful if the original base qualities have not been binned. Default bin values are 2,6,14,21,27,32,36

### ***Homopolymer Run Length Errors Statistics:***

--hpstats [file]	Collects counts on homopolymer run length errors (e.g. IONTorrent) and writes them to the named file at end of run. Default filename is indels.tsv. Charts can be produced from this file using the script IonTorrent.R IONTorrent.R [-i indels.tsv] [-r indelreport.pdf] For the XY charts to work Novoalign needs to parse XY coordinates from the read headers. This has been tested for Illumina MiSeq (CASAVA 1.8), IONTorrent and 454.
------------------	---

### ***Platform Specific Tuning:***

--tune [Default|HiSeqX|HiSeq|NextSeq|NOVASEQ|BGISEQ500|MGISEQ2000|IONTorrent-1|IONTorrent-2|V3-Defaults]

#### **Platform specific settings...**

Default	-g 40 -x 2 --matchReward 4 --softclip 40,25 -H 5 -t 0,3.0 --hlimit 8
HiSeqX	-g 40 -x 2 --matchReward 4 --softclip 50,30 -H 7 -t 0,1.0 --hlimit 9 -p 4,20
HiSeq	-g 40 -x 1 --matchReward 4 --softclip 45,30 -H 17 -t 0,2.5 --hlimit 9
NextSeq	-g 35 -x 1 --matchReward 4 --softclip 35,25 --trim3hp AG -H 22 -t 0,2.0 --hlimit 8
NOVASEQ	-g 40 -x 1 --matchReward 3 --softclip 55,20 -H 17 -t 0,3.0 --hlimit 8
BGISEQ500	-g 40 -x 1 --matchReward 4 --softclip 50,30 -H 12 -t 0,2.0 --hlimit 7 -k
MGISEQ2000	-g 40 -x 1 --matchReward 4 --softclip 50,25 -H 12 -t 0,1.5 --hlimit 8
IONTorrent-1	-g 100 -x 2 --matchReward 2 --softclip 100,25 -H 5 -t 0,3.5 --hlimit 9 -k
IONTorrent-2	-g 90 -x 2 --matchReward 2 --softclip 80,40 -H 15 -t 0,4.0 --hlimit 9
V3-Defaults	-g 40 -x 6 --softclip 0,0 -H off -t 16,4.5 --hlimit 9 -v 70 -r None --pechimera off

These settings were chosen to optimise F2 score using publicly available NA12878 WES datasets, Freebayes against GIAB High Confident SNPs.

## 4.2 Examples

`novoalign -f s_1_sequence.fq -d celegans.nix`

Aligns the reads in file `s_1_sequence.fq` against the indexed genome of *C.Elegans*.

`novoalign -a -m -f s_1_0001.fq -d hg36.nix`

Aligns a set of miRNA reads against the human genome. Adapter sequences are trimmed from the reads and an additional miRNA hairpin score is given for each alignment. Reports multiple alignments per read if they exist.

`novoalign -R 30 -r All -f s_1_sequence.fq -d hg36.nix`

Aligns a set of reads against indexed human genome, reporting multiple alignments per read. Any read with a score within 30 points of the best alignment will be reported.

`novoalign -f sim_l.fastq sim_r.fastq -d chrX.nix`

Aligns the paired files '`sim_l.fastq`' and '`sim_r.fastq`' against an index `chrX.nix`

## 4.3 Description

### 4.3.1 Base Qualities and Alignment Scores

Novoalign aligns reads against a reference genome using qualities and ambiguous nucleotide codes. The initial alignment process finds alignment locations in the indexed sequence that are possible sources of the read sequence. The alignment locations are scored using the Needleman-Wunsch algorithm with affine gap penalties and with position specific scoring derived from the read base qualities and any ambiguous codes in the reference sequence. User defined affine gap penalties are used for scoring insert/deletes.

Novoalign uses Needleman-Wunsch alignments with affine gap penalties, the gap opening penalty should be set to  $-10\log_{10}(P_{\text{gap}}) - G_{\text{extend}}$  where  $P_{\text{gap}}$  is the probability of an insertion deletion mutation vs the reference genome and  $G_{\text{extend}}$  is the gap extension penalty. Likewise the gap extend penalty can be set to  $-10\log_{10}(P_{\text{gap}2}/P_{\text{gap}1})$  where  $P_{\text{gap}1}$  is the probability of a single base indel and  $P_{\text{gap}2}$  is the probability of a 2 base insert/delete mutation. The default gap penalties were derived from the frequency of short insert/deletes in human genome resequencing projects.

Base quality values are used to calculate base penalties for the Needleman-Wunsch algorithm. The base qualities are converted to base probabilities and then to score penalties.

#### 4.3.1.1 PRB Quality to Score Conversion

The prb file has quality score  $Q(b, i)$  for each base,  $b$ , at each position,  $i$ , in the read. The quality

value is converted to a probability,  $Pr(b, i)$  and then to a penalty  $P(b, i)$ .

$$Pr(b, i) = \frac{10^{\frac{Q(b, i)}{10}}}{(1 + 10^{\frac{Q(b, i)}{10}})}$$

$$P(b, i) = -10 \times \log_{10}(Pr(b, i))$$

#### 4.3.1.2 Single Base Quality to Score Conversion

FASTQ reads and other read formats have a called base  $S(i)$  and single quality score  $Q(i)$  at each position,  $i$ , in the read. The quality value is converted to a probability,  $Pr(i)$  and then to a penalty  $P(S(i), i)$ .

Solexa

$$Pr(i) = \frac{10^{\frac{Q(i)}{10}}}{(1 + 10^{\frac{Q(i)}{10}})}$$

Fastq or Phred

$$Pr(i) = 1 - 10^{-\frac{Q(i)}{10}}$$

Alignment Penalty

$$P(S(i), i) = -10 \times \log_{10}(Pr(i))$$

$$P(b \in (\{A, C, G, T\} \setminus S(i)), i) = -10 \times \log_{10}((1 - Pr(i)) \div 3)$$

#### 4.3.1.3 Base Penalty Limit

For nucleotide alignments the penalties calculated above are further limited to a maximum of 30 at any base position.

#### 4.3.1.4 Base Quality to Penalty Table

The following table illustrates the conversion of base qualities to alignment score penalties. Other factors affecting penalties include ambiguous IUPAC codes in the reference and quality calibration.

**Note.** Very low quality bases can contribute to alignment score even if they match the reference! It is not possible in Novoalign to use the threshold parameter to control the number of mismatches allowed in the alignments. The threshold sets a lower limit on the probability that the aligned sequence could have generated the read.

Base Quality	Match Penalty	Mismatch Penalty
1	6	6
2	6	6
3	3	8

4	2	9
5	2	10
6	1	11
7	1	12
8	1	13
9	1	14
10	0	15
Q>10	0	Min(30, Q+5)

#### 4.3.1.5 Match Reward

The match reward is factored into the alignment score as additional penalty for inserted bases and is also used in soft clipping.

### 4.3.2 Posterior Alignment Probabilities and Mapping Quality

The posterior alignment probability calculation includes all the alignments found; the probability that the read came from a repeat masked region or from any regions coded in the reference genome as N's; and an allowance for a chance hit above the threshold based on the mutual information content of the read and the genome.

A posterior alignment probability,  $P(A_i | R, G)$  is calculated as:

$$P(A_i | R, G) = \frac{P(R | A_i, G)}{P(R | N, G) + \sum_i P(R | A_i, G)}$$

where  $P(R | N, G)$  is the probability of finding the read by chance in any masked reference sequence or any region of the reference sequence coded as N's, and where  $\sum_i$  is the sum over all the alignments found plus a factor for chance alignments calculated using the usable read and genome lengths.

The  $P(R | N, G)$  term allows for the fact that a fragment could have been sourced from portions of the genome that are not represented in the reference sequence. For instance in Human genome build 36 there is approximately 7% of sequence represented by large blocks of N's.

A mapping quality score is calculated as  $\min(70, -10\log_{10}(1 - P(A_i | R, G)))$ , where  $P(A_i | R, G)$  is the probability of the alignment given the read and the genome.

### 4.3.3 Mapping Quality and Alternative Scaffolds

The GRCh38 build of the human genome includes alternative scaffolds for some regions of the genome. Novoalign recognises sequences with FASTA headers that include an rg: tag and an rl:alt-scaffold tag, as in the following example, as alternative scaffolds, all other sequences are considered main scaffolds.

```
>chr4_KI270788v1_alt AC:KI270788.1 gi:568335894 LN:158965 rg:chr4:120164199-120317014 rl:alt-scaffold AS:GRCh38
```

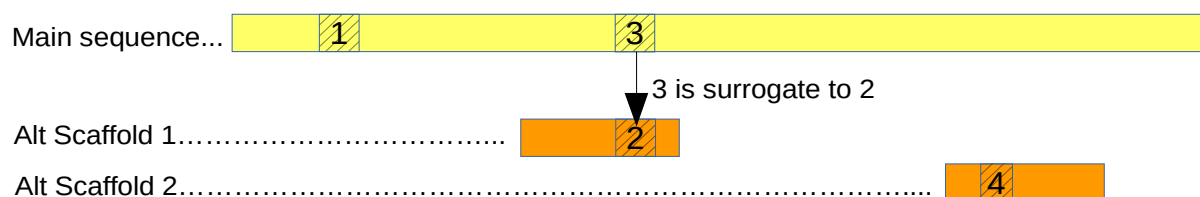
GRCh38 reference in this format can be downloaded from UCSC at <ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/analysisSet> file hg38.fullAnalysisSet.chroms.tar.gz includes the alt-scaffolds.

Terms	
ALT	Alternate scaffold sequences
REF	Main reference sequences
ALT-REF	A region on the main sequences that has one or more alternative scaffolds
Alternate mapping	Refers to any mapping on an ALT sequence or a ALT-REF with alternative scaffolds.
Surrogate mapping	Refers to any mapping that has an alternate mapping in ALT or ALT-REF with a better alignment score.
primary	Means secondary flag is not set

If the option --alt is specified on the command line the following rules apply...

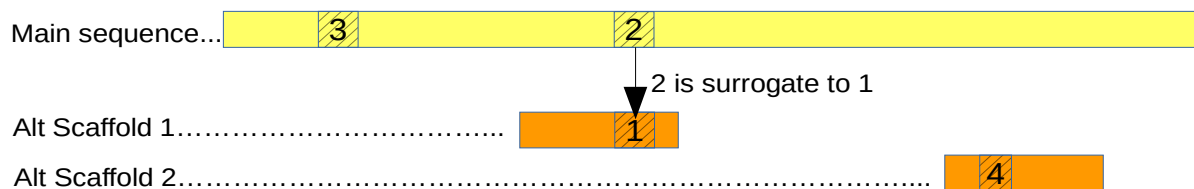
1. The best mapping (lowest alignment score) is primary. When several mappings have the same low score we randomly select the primary with preference to a main sequence mapping.
2. Mapping quality (posterior alignment probability) calculations for a mapping exclude any supplementary alignments.
3. Any mapping that is a surrogate to a lower scoring mapping is flagged as supplementary.
4. Supplementary mappings do not count toward the reporting limit.
5. Supplementary mappings that are surrogate to a primary mapping are also primary.
6. For proper pairs and single end reads the best REF mapping is primary

As examples consider in the following diagrams.



*Illustration 1: We have 4 mappings for a read with alignment 1 having the best score, 2 the second best etc. Alt\_scaffold 1 is an alternate for the portion of the main chromosome that contains mapping 3. The MAPQ for alignment 1 is calculated using mappings 1,2 & 4. Mapping 1 is primary and mapping 3 is supplementary, 2&4 are secondary.*





*Illustration 2: We have 4 mappings for a read with alignment 1 having the best score, 2 the second best etc. Alt scaffold 1 is an alternate for the portion of the main chromosome that contains mapping 2. The MAPQ for alignment 1 is calculated using mappings 1,3 & 4. The MAPQ for alignment 2 is calculated using mappings 2,3 & 4. Mapping 1 is primary and mapping 2 is supplementary.*

#### 4.3.3.1 Novoalign / BWA MEM Differences

1. Supplementary flag is set differently.
  - BWA sets supplementary on all mappings to ALT when there is a mapping to REF
  - For Novoalign the best mapping is not supplementary even if on ALT. Surrogate mappings are supplementary.
2. MAPQ is calculated differently
  - BWA the MAPQ of a REF hit is computed across REF hits only. The MAPQ of an ALT hit is computed across all hits.
  - In Novoalign the MAPQ is computed excluding surrogate mappings. MAPQ of surrogate mappings exclude any alternate mappings for the same ALT-REF region.
3. Novoalign adds a tag ZA:i: with a new MAPQ calculated from all mappings.
4. Primary flag is set differently. In Novoalign primary flag is set for the best alignment, all surrogates to the best alignment and on the best REF mapping (except for improper pairs). It is possible to be primary and supplementary.
5. What mappings get reported is likely different. Novoalign's reporting options for multi-mapped reads with ALT mappings have changed from earlier releases. For purpose of counting reported alignments and identifying multi-mapped reads surrogate mappings are not counted.

#### 4.3.4 Quality Trimming

The quality trimming option -H enable trimming of maximal number of 3' bases with average quality below the specified limit.

Format:

-H limit margin

Starting at 3' most base of each read we calculate the (base quality – limit), keeping a running sum of this value. If the running sum exceeds the margin, or we reach the 5' end of the read, then bases from the minimum value to the 3' end of the read are trimmed. The margin just serves to stop the scan for minimum value once the sum has gone high enough and may save a small amount of CPU time. The default is 30.

This is illustrated in the examples (Illumina HiSeqX) below. The green dot marks the start of the trimming, and the red arrow the end of the scan, for each example.



## 4.3.5 Adapter Trimming

### 4.3.5.1 Single End Reads - miRNA

Adapter trimming does a gapped global alignment of the adapter against the read and then trims the read from the start of the optimum alignment.

A few details:

1. The read and base qualities are first converted to a weight matrix where each base will score  $\max(30, -10\log(P))$  where  $P$  is probability of the base. This results in a match scoring 0 and a mismatch at high quality base position scoring 30
2. During adapter trimming we subtract 7 from the weights so at a high quality base position a match scores -7 and a mismatch 23.
3. If the optimum alignment scores  $\leq -7$  it is trimmed.
4. There are no penalties for unmatched letters at the beginning of the read or at the end of the adapter.

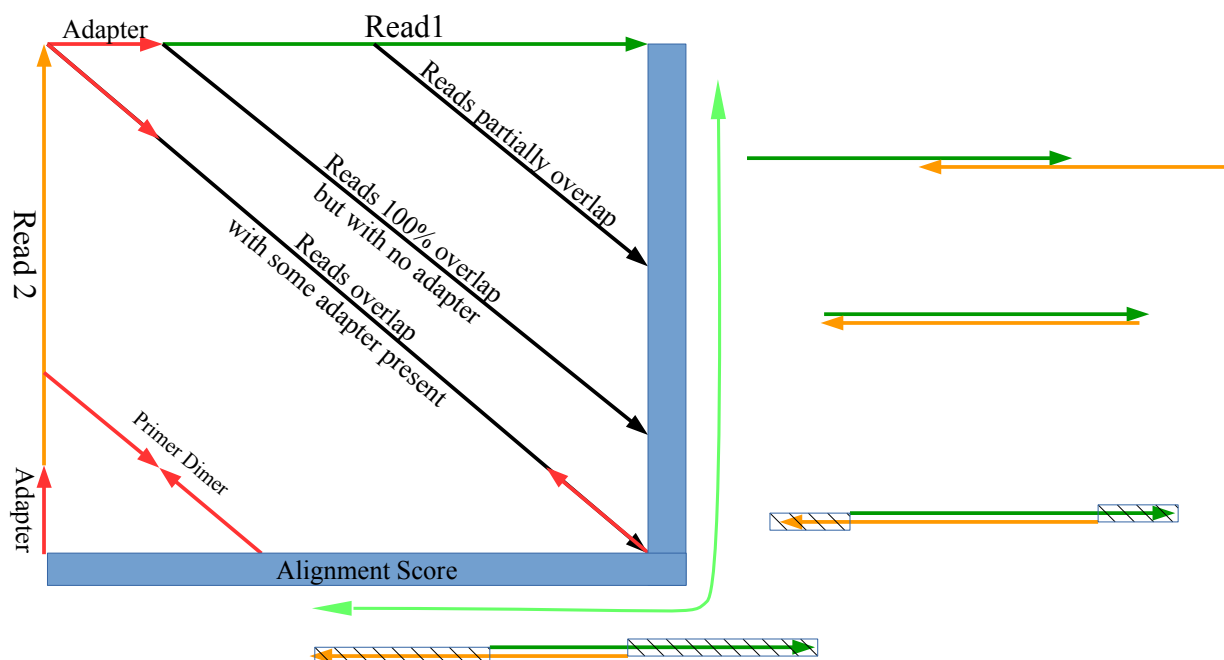
### 4.3.5.2 Paired End Reads – Short Fragments

If a DNA fragments is shorter than the read length then both reads of the pair will have extended into adapter or primer sequence and unless trimmed off will be used in alignment.

If there are only a few bases of adapter the read may still align but with some mismatches or indels in the adapter portion of the alignment. This contributes to SNP noise and reduced consensus quality so it's always best to trim them.

When there is more than a few bases of adapter the read is unlikely to align which isn't a problem except that there has been an attempt to align it that will have tried to align with possible 8 mismatches and up to 7 indels. This attempt to align the read with so many mismatches can consume considerable CPU time so it's desirable to identify these reads before aligning them.

Novoalign identifies short fragments by aligning the two reads of a pair against each other to detect overlap and adapter sequence. If overlap is detected then any adapter is trimmed from the two reads.



*Drawing 3: Dynamic Programming alignment of two paired-end reads with in silico prepended the first 20bp of the adapter sequence. High scoring diagonal alignments identify the amount of overlap and adapter sequence present in the read. False positive rate is low as reads must be complementary and align to the adapter to get a good score.*

You can use -a "" "" to detect short fragments with an unknown adapter sequence but this can't detect adapter dimers and to avoid false positives the 3' of the reads need to overlap by more than 25%.

When you know the adapter sequence specify more than 10bp if possible. If 20 or less bp are specified primer dimers may not be detected.

### 4.3.6 Amplicon Clipping

For targeted amplicon sequencing it is often desirable to exclude the amplicon primer sequence from the variant calling process. To facilitate this Novoalign includes an option to soft clip primer sequences from read alignments.

Reads are aligned using all the bases including the primer and then, if the read alignment maps to an amplicon, the primer bases are soft clipped from the alignments. A read is considered to match to a primer if the mapping location falls within the primer region. In normal operation if the mapping is even 1bp before the primer it will not be detected as a match and will not be clipped. This can be changed by allowing a few bp deviation in mapping location using the delta option.

For paired end reads the reads 5' alignment locations are checked against primer locations and trimmed accordingly. Then a check is done for each read to see if read 3' alignment overlaps with the same primer trimmed from it's mates 5' end and if so the 3' end is trimmed. This allows for amplicons where the insert is shorter than the read.

In paired end mode we allow the two reads of the pair to align to primers of different amplicons.

In single end mode the read 5' is checked for alignment to a primer and if so it's soft clipped. The 3' alignment location is then checked for alignment to the other primer of the same amplicon and if so it is soft clipped. This allows for reads shorter than the insert length.

Amplicon soft clipping is enabled by the option...

--amplicons *amplicons.BED [delta] [output.bed]*

#### Bed File Format

chrom      Name of the chromosome  
 chromStart   Start position of the amplicon (includes primer bases)  
 chromEnd    End position of the amplicon (This is one bp passed the end of the primer)  
 name        Amplicon name if any  
 score        ignored on input, set to count of matching read pairs on the output BED file.  
 strand       + or -, ignored for now.  
 thickStart   Start of amplicon excluding primer  
 thickEnd    Start of the second primer (i.e. One bp passed the amplified region)  
 itemRgb     ignored

**Note.** BED files use zero based coordinates and SAM files are 1 based. Also, the chromEnd and thickEnd are open which means they are not included in the amplicon. If you find your primers are not being soft clipped then check that the amplicon coordinates are not off by 1.

#### Example

```
chr2 29083861 29084059 AMP.1 100 - 29083881 29084039
chr2 29085075 29085273 AMP.2 100 - 29085095 29085254
chr2 29089969 29090233 AMP.3 100 - 29089989 29090214
chr2 29091056 29091241 AMP.4 100 - 29091076 29091220
```

At the end of the alignment process the counts of amplicon clipping events are printed to the Novoalign log.

e.g.

#	Amplicon	Count	SE5	SE3
#	AMP.1	371	0	0
#	AMP.2	190	0	0

There are three counters, first is the number of hits where both reads of pair aligned to primers of the same amplicon. The next two counts are where read1 & read2 of pair aligned to different amplicons or perhaps one read of the pair failed to align to an amplicon.

There is an optional SAM tag, ZP, to report the amplicon name in the case where a read pair has matched an amplicon. Use option --tags ZP to enable.

### 4.3.7 Read Quality

Reads with too many low quality base positions will not be aligned. This is controlled by the `-l` options and effectively sets the minimum length, or minimum number of high quality base positions in order for an alignment to be attempted. The read length calculation uses base qualities to calculate the information content of the read.

Homopolymer reads are also deemed low quality and not aligned. These are fairly frequent in real data and are possibly the result of dust on slides.

### 4.3.8 Reads with Multiple Alignments

There are times that reads will align to multiple locations with very similar alignment scores. Situations where this might occur are reads originating from repeats and the alignment of very short reads such as small RNA.

Depending on the users project and objectives, reads and alignments may be or not be of interest.

Every read will have multiple alignment locations however the alignment score could be very different, so for detection of repeats novoalign programs use the difference in score between the best alignment and the rest of the alignments. This score difference is set by the `'-R99'` option and defaults to 5 which corresponds to the best alignment being approximately 3 times more probable than the next best alignment. For example, two alignments with probabilities 0.7 (score 1) and 0.3 (score = 5) would be considered as multiple alignments to the read. Two alignments with probabilities 0.8 (Score 0) and 0.2 ( score 7) would be treated as a unique alignment to the location with the higher probability.

Having identified a read as having multiple alignment locations we then have several options for reporting.

Option	Description
None	No alignments will be reported. The read will be reported as a status R with a count of the number of alignments. No alignment locations will be reported.
Random	A single alignment location is randomly chosen from amongst the alignment results. The choice is made using posterior alignment probabilities.
All	All alignment locations are reported. Note, that this is all alignments with a score within 5 points of the best alignment unless you use the <code>-R99</code> option to extend the range.
Exhaustive	This option bypasses the iterative alignment process and the normal repeat alignment detection. It finds all alignments with a score no worse than the threshold ( <code>-t 99</code> option) and reports all the locations.

### 4.3.9 Sequence file formats

Read files are introduced using the `-f` options. Novoalign examines the file name and the first few

lines of each file to determine the file format.

Licensed versions of Novoalign will also process read files compressed with gzip.

Format	File Names	Description and detection method
FASTA	*.fa *.fna *.fasta	<p>Standard FASTA format input file can be used. This file type is recognised by the name matching *.fa, *.fna , or *.fasta or by the first line starting with a '&gt;' character. e.g.</p> <pre>&gt;sequence_0 GATGTCCTCAGTATGAGAAAGAGGCAGGTTCTGGG &gt;sequence_1 ACACGCAGCGCCGCGCATGCTTGCGCCGCCACTCCA &gt;sequence_2 ACCTGCGCTCTGCCCTGAAACCACTGTTGGCTTGAG</pre> <p>Example:  <code>novoalign -f reads.fa -d celegans</code></p>
.FASTA & Quality	as above with *.qual	<p>Fasta file are detected and then the folder is checked for a quality file. If Novoalign detects a fasta format read file it looks for a matching *.qual file in the same folder. If found then it will be used for base qualities.</p> <pre>&gt;sequence_0 40 40 40 40 40 40 40 40 40 40 40 40 14 40 40 40 40 40 40 40 40 25 40 40 40 40 40 40 5 40 8 9 21 40 4 &gt;sequence_1 40 19 7 22 4 40 8 40 40 40 9 40 28 40 40 40 17 31 11 40 32 24 4 9 14 10 36 16 40 9 2 8 6 16 3 3</pre>
Sanger FASTQ	*.fastq	<p>Sanger format FASTQ files are recognised by the file name matching *.fastq. Quality scores are from ASCII code – 33.</p> <p>For non-standard file names this format is detected by an '@' character starting the first line and by a test on the quality codes of the first read. Sanger fastq files are automatically detected as the ASCII coded qualities are lower than for a Solexa format FASTQ file.</p> <p>Example:  <code>novoalign -f reads.fastq -d celegans</code></p>



Solexa  
FASTQ  
and  
Illumina  
FASTQ

\*\_sequence.txt

Files produced by Illumina pipeline with Solexa variant of the FASTQ format. Solexa quality scores are ASCII letter code – 64; See Gerald documentation for a full description. These files are named like s\_lane\_sequence.txt and recognised by matching the file name against s\*\_sequence.txt.

For non-standard file names this format is detected by an '@' character starting the first line and by a test on the quality codes of the first read. Solexa fastq files are automatically detected as the ASCII coded qualities are higher than for a Sanger format FASTQ file.

Starting from Version 1.3 of the Illumina Casava Pipeline the coding of quality values was changed to the Phred scale. If you are using Pipeline 1.3 you may need to add the option -F ILMFQ. This option will treat quality codes as being coded as  $-10\log_{10}(\text{Perr}) + '@'$ . The old Solexa format is the default for \_sequence.txt files and interprets quality values according to formula  $-10\log_{10}(P/(1-P)) + '@'$

Illumina  
Casava 1.8  
FASTQ

\*\_sequence.txt

New format introduced in Casava V1.8 these base qualities are now coded in Sanger format. The header also includes an 'is\_filtered' field that is set to 'Y' if the base caller has flagged the read as low quality (more details below). By default low quality reads will be skipped. Refer to -F command line option for further options.

```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
BBBBCCCC?<A?BC?7@@???????DBBA!!!!A##
```

From [seqanswers.com](http://seqanswers.com):

The Illumina is\_filtered flag is based solely on the relative intensity of the fluorescent signals. There are two methods Illumina uses to calculate relative intensities called Chastity and Purity. Chastity is defined as the ratio of the intensity of the most intense base for a cluster divided by the sum of the most intense plus the second most intense signal. Purity is defined as the ratio of the most intense signal divided by the sum of all four fluorescent signals. The default parameter used by GERALD when filtering reads is  $\text{CHASTITY} \geq 0.6$ . Stated another way (after doing a little algebra) the most intense signal must be at least 1.5x higher than the second most intense signal. Also, filter passing is only based on the signals over the first 12 cycles. I am not sure whether this means that the value must be  $\geq 0.6$  for each of those 12 cycles or that average is  $\geq 0.6$ .

This filter is designed to detect polyclonal clusters.





Reads are extracted from the BAM file.  
BAM aligns both single & paired reads  
BAMPE only aligns paired reads, single end reads are skipped.  
BAMSE treats each BAM record as a single end read.  
Secondary alignments are skipped.  
The BAM file should be in an order such that paired reads are adjacent in the file.

Tab separated single line read format with Sanger quality scores. Each line has these three (for single end reads) or five tab-delimited fields:

1. Read Name
2. Sequence of first fragment
3. Qualities of first fragment
4. Sequence of second fragment
5. Qualities of second fragment

Single & Paired end reads can be mixed.

### 4.3.10 Output Formats

Three output formats are provided.

1. Native
2. Extended Native
3. Pairwise
4. SAM

#### 4.3.10.1 Native Report Format

The native format is designed to be compact, giving essential information necessary for downstream processing. This is default report format.

```
# novoalign (1.0) - short read aligner with qualities.
# (C) 2008 NovoCraft
# Licensed for evaluation and educational Use Only
# novoalign -d ssuis -f ../../s_8_0100/s_8_0100.fa -q ../../s_8_0100/s_8_0100.qual
# Index Build Version: 1.0
# Hash length: 11
# Step size: 1
# Interpreting input files as FASTA with Phred quality file.
>I8_100_293_551 S CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACC IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII NM
>I8_100_880_947 S TTATTATCTTTATTGACGTACCTCTAGAAGACCCAA IIIIIIIIIIIIIIIIIIIIIIIIIIII;>1 U
    0   150 >SsuIs 420732 R . . .
>I8_100_975_684 S AGTAGACACCTGGTGAACGAACCAACTGAGAAACGA IIIIIIIIIIIIIIIIIIIIIIIII-EII)IIIIIG U
    1   150 >SsuIs 111343 R . . .
>I8_100_874_727 S GTGAAAGCCAGCGTCTTTAGGCCTGGGTGGTGGTG IIIIIIIIIIIIIIIIIIIIIIIII%IIIII,59 R
    4
>I8_100_244_639 S AACATAATTAGACAGAATATAAGATATGACTAATTC IIIIIIIIIIIIIIIIIIIIIIIIIII9H2)I U
    1   150 >SsuIs 1364843 R . . .
>I8_100_492_8 S ANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN I"""""""""""""""""""""""""""""" QC
>I8_100_515_741 S GGAAATCACGGAGCAGGAGTTTCGTGAGCTTCGCCG IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII/I U
    49  141 >SsuIs 429042 F . . . . . 35G>C 36A>G
>I8_100_510_804 S AACCAGACGTTGCTTCGTCTACAATCACAATACCCG IIIIIIIIIIIIIIIIIIIIC9II='II$II&I,&H89+0 U
    54  117 >SsuIs 1499130 R . . . . . 4C>G 9T>G 15T>G
>I8_100_188_601 S ACTACGTTACAGAAAAATCTAGCCTTTGTACTAGAC IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII/II% U
    9   150 >SsuIs 145620 R . . . . . 1T>G
>I8_100_63_601 S AGCGGCAGGGCTTGTTCCAGCTAAGCTCCGATTTT IIIIIIIIIIIIII'IIIIIIIIHII%A%,I&IIII U
    114 57 >SsuIs 1997459 R . . . . . 8T>G 9T>A 11T>C 22T>A 27T>C
```

```
>I8_100_331_271 S GGATTATGTGAAACAACATGCTGATGCACCGCTTAA IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII,II@& U
18 150 >Ssuis 1883394 R . . . 5T>G
>I8_100_408_934 S ATGATATTAGGTCCTATCTTACTTTTCTCAACCAAC IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII= U
0 150 >Ssuis 580585 R . . .
>I8_100_269_390 S GTGTTCCCAAACCTGCTGCAGGGATAACGGCTTTT IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII8 U
28 150 >Ssuis 1853977 R . . . 1G>A

...
>I8_100_768_102 S AAACACATGGTGTATNAAACTCGCGACGTAGTCAT )%)II"&I$,I9I)")"I'*7IGI9"2'IE$$I&%$ NM
>I8_100_582_231 S TAAGCAAAAAACATAATTCCAGGATATGCAACCAGT '%"&#&)$%$$"$#$#%'%$%##"$)%#%$#$" QC
>I8_100_240_200 S AATAAAGCCTAAACAATGGACAACAACTACACAC :%$%/$#&$&$&$&$"$#$&#%'%$%$#$#$&'& QC
# Paired Reads: 5000
# Proper Pairs: 3446 (68.9%)
# Read Sequences: 10000
# Unique Alignment: 8015 (80.2%)
# Multi Mapped: 0 ( 0.0%)
# No Mapping Found: 1985 (19.9%)
# Elapsed Time: 0.735 (secs.)
# CPU Time: 1.24 (secs.)
```

Normally a read is printed on one line with a series of tab delimited fields. The fields are :-

Field	Description												
Read Header	The fasta or fastq header of the read sequence.												
S, L or R	S indicate this is an alignment for single ended read. For paired end reads L indicates the read is from the first file. R indicates the read is from the second file.												
Read Sequence	The read sequence.												
Base Qualities	Standard (Sanger) Fastq format base qualities, empty for fasta input unless using quality calibration. If quality calibration is used these are calibrated qualities.												
5' trim count	Count of bp trimmed from the 5' end of a read. Refer -5 command line option. Only present in Extended Native format												
3' trim count	Count of bp trimmed from the 3' end of a read. Refer -a & -s command line options. Only present in Extended Native format												
Status	<table> <tr> <th>Status</th><th>Meaning</th></tr> <tr> <td>U</td><td>A single alignment with this score was found.</td></tr> <tr> <td>R</td><td>Multiple alignments with similar score were found.</td></tr> <tr> <td>QC</td><td>The read was not aligned as it bases qualities were too low or it was a homopolymer read.</td></tr> <tr> <td>NM</td><td>No alignment was found.</td></tr> <tr> <td>QL</td><td>An alignment was found but it was below the quality threshold.</td></tr> </table>	Status	Meaning	U	A single alignment with this score was found.	R	Multiple alignments with similar score were found.	QC	The read was not aligned as it bases qualities were too low or it was a homopolymer read.	NM	No alignment was found.	QL	An alignment was found but it was below the quality threshold.
Status	Meaning												
U	A single alignment with this score was found.												
R	Multiple alignments with similar score were found.												
QC	The read was not aligned as it bases qualities were too low or it was a homopolymer read.												
NM	No alignment was found.												
QL	An alignment was found but it was below the quality threshold.												
Alignment Score	This is the Phred format alignment score $-10\log_{10}(P(R A_i))$ . For status of 'R' and when not report alignment locations for repeats, this field becomes the number of alignments to the read. For paired end the alignment score includes the fragment length penalty.												
Alignment Quality	This is the Phred format alignment quality score $-10\log_{10}(1 - P(A_i R, G))$ using Sanger fastq coding method.												

<i>Proper pair flag</i>	A value of 1 indicates that the read pair was aligned as a proper pair. Only present in extended native format.
miRNA score	Alignment score for adjacent opposite strand alignment. Optional, only included in miRNA mode.
Aligned Sequence	The fasta header of the aligned sequence. This is truncated at first space.
Aligned Offset	The 1-based position of the alignment in the sequence.
Strand	F/R Indicator of alignment direction.
Pair Sequence	The fasta header of the sequence the reads pair was aligned to. For single ended reads, or pairs where both ends aligned to the same sequence, this field is set to '!'. If a paired alignment that fits the fragment length distribution is not found and we are reporting two individual alignments for the pair then the pair alignment location is only reported if both alignments have an alignment quality > 10.
Pair Offset	The 1-based position of the alignment to the pair of this read. For single ended reads this field is a '!'. In miRNA mode we report the alignment location for adjacent opposite strand alignment.
Pair Strand	F/R Indicator of alignment direction of the pair of this read. '.' for single ended reads.
Mismatches	A list of base indels, mismatches and bases inserted or deleted. Format is 'offset'refbase>'readbase' where the offset is 1 based position of difference relative to the 'Aligned Offset'.  <b>Note.</b> Offset of mismatches are relative to the alignment location. They are not the location of the mismatches in the read. This distinction is important when the alignment contains indels and/or is soft clipped back to the best local alignment. Inserts are in format 'offset'+ 'insertedbases' and deletes in format 'offset'- 'refbase' The mismatch list is space delimited. A mismatch is only reported if the probability of the base is less than 0.16. For fastq files this corresponds to a $P_{err} \approx 0.5$ When using soft clipping the number of bases soft clipped from the 5' (as aligned) end of the alignment is reported using format 0x'n', and for 3' end as 'offset'x'n' where n is the number of bases soft clipped.

#### 4.3.10.2 Paired End Native Report Format

This example is for native format with good pairs found. The alignment score for one of the reads in the pair will include the fragment length penalty. The quality score is based on the posterior fragment alignment probability.

```
# novoalign (2.0) - short read aligner with qualities.
# (C) 2008 NovoCraft
# Licensed for evaluation and educational Use Only
# novoalign -d ssuis -f ../../simlft/s_1_sequence.txt
```

```

.././simrgt/s_1_sequence.txt
# Index Build Version: 1.0
# Hash length: 11
# Step size: 1
@Ssuis_633667_633825_0/1 L GCTCAATGACTATCCGCAGATTGAGGGGTTTCTGCT
IIIIIIIIIIIIIIIIII,IIIII(,;!%$3C;I>!!U 51 150>Ssuis 633790 R . 633667 F
@Ssuis_633667_633825_0/2 R GTCTGACTCATGGCTGTGCGAATGGCTTCTTCCCTAIIIIIIIIIIII-
IIIIIIIIIIIIIIIIII0%!!U 16 150>Ssuis 633667 F . 633790 R
@Ssuis_1657428_1657600_1/1 L AGTACGTGTCAATATCGTCCACTCTGCAGGTGGTCC
IIIIIIII+IIIIIIIIIIIIIIIIIIII+CB-4%7U 42 150>Ssuis 1657565 R .
1657428 F 2C>G 7A>C
@Ssuis_1657428_1657600_1/2 R TGTAATGATGCTGTGAAGACGTACTTCAACATCAT
IIIIIIIIIIIIIIIIIIIBIIII6I<)IIIII)I+7(U 3 150>Ssuis 1657428 F .
1657565 R
@Ssuis_973563_973724_f/1 L TTACCAAGCGTGGTAATCCCTACGCTAGAAAGATTC
IIIIIIIIICIIIIIIIIIIIIII'%'$2III-II-2R 2
@Ssuis_973563_973724_f/2 R TGGCACCAATCGTGTGCAGCTTCGTTGAAGTCGTTT III%!!%
+IIIIIIIIIIIIIIIIII+IIIIII,II R 2
...
# Paired Reads: 5000
# Proper Pairs: 3446 (68.9%)
# Read Sequences: 10000
# Unique Alignment: 8015 (80.2%)
# Multi Mapped: 0 ( 0.0%)
# No Mapping Found: 1985 (19.9%)
# Elapsed Time: 0.735 (secs.)
# CPU Time: 1.24 (secs.)
# Fragment Length Distribution
# From To Count
# 45 59 1
# 60 74 2
# 75 89 4
# 90 104 5
# 105 119 23
# 120 134 34
# 135 149 72
# 150 164 101
# 165 179 163
# 180 194 208
# 195 209 323
# 210 224 353
# 225 239 405
# 240 254 398
# 255 269 356
# 270 284 318
# 285 299 256
# 300 314 170
# 315 329 114
# 330 344 73
# 345 359 37
# 360 374 21
# 375 389 6
# 390 404 1
# Mean 240, Std Dev 51.1
# Done at Fri Jun 6 14:30:37 2014

```

This example is for native format when a good pair was not found. In this case both alignments

were on different chromosomes. The quality values reflect the quality of the individual end alignments.

```
@SLXA-EAS1_34_FC4751_R1_1_1_53_21 L
TTGATGGATCAATTGTAGTTGCCTGCAATAAGAGG ??????????????????????:????9+2$
U 23 150 >III 7197040 R >IV
11532213 F 3G>T
@SLXA-EAS1_34_FC4751_R1_1_1_53_21 R AATTGGAAGAGGACAGAAGAGATGA
=====8==+ U 1 93 >IV 11532213
F >III 7197040 R
@SLXA-EAS1_34_FC4751_R1_1_2_993_712 L
GTGCCTACCATTTGTGATTCGACTATATACGCGCTC ??????8?8?????????5?09?5?7?*(&&7%7,
U 6 150 >IV 5943661 F >I 4229259 R
@SLXA-EAS1_34_FC4751_R1_1_2_993_712 R GGGAAAAGGTGCCAAAAGTATAGA
<<<1<<<<<<1<-//4-<31<3<- U 0 94 >I 4229259 R
>IV 5943661 F
```

This example is for native format with multiple alignments to a read and using -r All option.

```
>8_100_1_16 L TTACCAAGCGTGGTAATCCCTACGCTAGAAAGATTC IIIIIIIICIIIIIIIIIIIIII%'%
$2III-II-2 R 6 3 >Streptococcus_suis 973563 F . 973689 R
>8_100_1_16 R TGGCBDAATCGTGTGCAGCTTCGTTGAAGTCGTTT III%"#%
+IIIIIIIIIIIIIIIIII+IIIIII,II R 41 3 >Streptococcus_suis 973689 R . 973563
F
>8_100_1_16 L TTACCAAGCGTGGTAATCCCTACGCTAGAAAGATTC IIIIIIIICIIIIIIIIIIIIII%'%
$2III-II-2 R 6 3 >Streptococcus_suis 1717310 R . 1717184 F
>8_100_1_16 R TGGCBDAATCGTGTGCAGCTTCGTTGAAGTCGTTT III%"#%
+IIIIIIIIIIIIIIIIII+IIIIII,II R 41 3 >Streptococcus_suis 1717184 F .
1717310 R
```

### 4.3.10.3 SAM/BAM Report Format

SAM report format is for use with SAMtools, just add the option -oSAM to the command line.

The report format is documented as part of SAM/BAM specification at

<http://samtools.sourceforge.net/>

The standard SAM tags Novoalign can add to alignments are...

Tag	Default	Description
AM	On	The smallest template-independent mapping quality of other segments in the read. Only for multi-template reads.
AS	On	Alignment score generated by Novoalign.
BC	Off	Extract barcode from Illumina header and add BC:Z: tag to SAM records. e.g. @EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG adds BC:Z:ATCACG to SAM records.
CC	On	Reference name of the next hit; '=' for the same chromosome. Only present if read has multi-mappings reported
CP	On	Leftmost coordinate of the next hit. Only present if read has multi-mappings reported
LB	Off	Library. This is extracted from the LB tag of the @RG record and is redundant.
HI	On	Query hit index, indicating the alignment record is the i <sup>th</sup> one stored in SAM. Only present if there is more than one alignment reported for the read.

IH	On	Number of stored alignments in SAM that contains the query in the current record. Only present if there is more than one alignment reported for the read.
MC	Off	For paired end reads the CIGAR of mate to this read.
MD	On	String for mismatching positions. Regex : [0-9]+((([A-Z]) ^[A-Z])[0-9]+)*6
NH	On	Number of reported alignments that contains the query in the current record. Only present if there is more than one alignment reported for the read. Note. In our interpretation of the SAM specification we have taken this as the count of alignments that would be reported if there wasn't a limit imposed by the -r option, so this is the number of alignments found with the quality range (-R) or threshold limits for -r Exhaustive. The IH tag is the number of alignments that were stored in the SAM file.
NM	On	Edit distance to the reference, including ambiguous bases but excluding clipping
OQ	Off	Original base quality. Useful with quality calibration (-k option)
PG	On	Program. Matches the @PG tag which is automatically added by Novoalign*
PQ	On	Phred likelihood of the template, conditional on both the mapping being correct. Only for multi-template reads.
PU	Off	Platform unit. This is extracted from the PU tag of the @RG record and is redundant.
RG	On	Read group. Present if an @RG record was entered with the -o SAM option
SM	On	Template-independent mapping quality. i.e. The mapping quality if mapped as a single end read. Only present for multi-template reads. Whilst Novoalign attempts to ensure this value is approximately correct there are no guarantees to its accuracy.
UQ	On	Phred likelihood of the segment, conditional on the mapping being correct
YH	Off	Hard Clipped Bases. A ' ' separates 5' & 3' hard clipped bases.
YQ	Off	Hard Clipped Base Qualities

Novoalign also adds several custom tags...

Tag	Type	Default	Description						
XM	i	Off	Count of mismatches in the alignment (excludes indels)						
XO	i	Off	Count of gap opens in the alignment						
YH	Z	Off	Hard Clipped bases. 5' & 3' bases are separated by a ' '. Not present if there is no hard clipping.						
YQ	Z	Off	Qualities for hard clipped bases.						
ZA	i	Enabled if reference has alt-scaffolds	MAPQ of read or pair using all mappings. Only present if the read had mappings to alternate scaffolds. Down stream programs might use this to identify which alternative scaffolds are present.						
ZB	Z	On	For Bi-Seq alignments indicates which index was used to align the read. This can be used to separate alignments by source strand of the DNA fragment.						
			<table><tr><th>Value</th><th>Meaning</th></tr><tr><td>CT</td><td>The C/T index was used for alignment. This means the fragment was from the 5' -3' strand of the chromosome.</td></tr><tr><td>GA</td><td>The G/A index was used. This means the fragment was from the 3'-5' strand of the chromosome.</td></tr></table>	Value	Meaning	CT	The C/T index was used for alignment. This means the fragment was from the 5' -3' strand of the chromosome.	GA	The G/A index was used. This means the fragment was from the 3'-5' strand of the chromosome.
Value	Meaning								
CT	The C/T index was used for alignment. This means the fragment was from the 5' -3' strand of the chromosome.								
GA	The G/A index was used. This means the fragment was from the 3'-5' strand of the chromosome.								
ZH	i	if -m On else Off	Hairpin score for miRNA alignment (-m option)						
ZL	i	if -m On else Off	In miRNA mode (-m option) this is the alignment location for adjacent opposite strand alignment.						

ZO Z On

Indicates long or short insert fragment for mate pair alignments when short insert has been enabled.

Value	Meaning
'+-'	Indicates pair was aligned as a short insert fragment.
'-+'	Pair was aligned as a long insert fragment.

This tag is only present for Illumina mate pairs when a short fragment length size has been specified with the -i option and reads are aligned as a proper pair .

ZP Z Off

Amplicon name. Use in conjunction with -amplicons option. Only present if both reads of pair mapped to the same amplicon. Counts of alignments with specific amplicon name may differ from Novoalign log as it includes reads where only one read of the pair mapped.

ZQ f if -q On else Off

MAPQ in float format with decimal places specified by -q option

ZS Z On

Novoalign alignment status. Not present for unique alignments.

Status	Meaning
NM	No alignment was found.
QC	The read was not aligned as it bases qualities were too low, it was a homopolymer read, or it failed the polyclonal filter check.
R	Multiple alignments with similar score were found.

Z3 i Off

3' mapping location. Only reported if option --3Prime is used.

SAM tags can be enabled or disabled using the --tags option, ALL operates on every tag. A minus '-' after the tag name disables the tag.

Examples

```
novalign --tags Z3 LB- PU- ...
```

```
novalign --tags ALL- ...
```

When using SAM report format the run headers and statistics normally output as part of Native format reports are written to stderr.

```
# novalign (V3.02.05 - Build Apr 16 2014 @ 12:35:03 - A short read aligner with qualities.
# (C) 2008,2009,2010,2011 NovoCraft Technologies Sdn Bhd.
# License file: /home/sparks/bin/novalign.lic
# Licensed to Novocraft Technologies Sdn Bhd (Internal Use Only)
# novalign -o Sy -d cp.nix -f ../data/chrend/se.ilm.bwa.read1.fastq
../data/chrend/se.ilm.bwa.read2.fastq -o SAM -o FullNW
# Starting at Wed Apr 16 12:35:37 2014
# Interpreting input files as Sanger FASTQ.
# Index Build Version: 3.2
# Hash length: 6
# Step size: 1
# Paired Reads: 5000
# Proper Pairs: 3446 (68.9%)
# Read Sequences: 10000
# Unique Alignment: 8015 (80.2%)
# Multi Mapped: 0 ( 0.0%)
# No Mapping Found: 1985 (19.9%)
# Elapsed Time: 0.735 (secs.)
# CPU Time: 1.24 (secs.)
```

#### 4.3.10.3.1 @SQ M5 tags

The SAM/BAM specification allows an optional M5 tag that has the MD5 checksum of the

reference sequence. Some downstream analysis programs require that this is present and others may check that the M5 tag value matches the supplied reference sequence.

There are several ways to get the M5 tags onto the @SQ records

1. If they are present on the fasta headers of the reference sequences used by NovoindeX they will be passed through to alignment via the sequence headers.
2. You can use Novoutil **fastaAddM5** function to add M5 and LN tags to your fasta headers.
3. NovoindeX can generate the M5 tags using the -m5 option. This will generate tags and can replace any existing tag with the generated value.
4. If Novoalign doesn't find the M5 tags on the sequence headers in the index it will generate the M5 tags based on the sequence in the index. This can be disabled with the option **--addM5 off**

If...

I don't want M5 tags!

1. Make sure your fasta reference doesn't have M5 tags.
2. Run novoindeX without the -5 option.
3. Add **--addM5 off** option to Novoalign command line

I'm adding IUPAC codes to my reference but I need M5's to match original reference!

1. If your original fasta reference sequences do not have M5 tags they can be added using **novoutil fastaAddM5** function
2. Use **novoutil IUPAC** to encode desired SNPs as IUPAC ambiguous codes
3. Run novoindeX without the -5 option.
4. Novoalign will add the M5 tags from the original fasta to the @SQ records

I want M5 tags to match the reference sequence used in novoindeX but M5 is not on my fasta headers!

1. Add option **-5** to the novoindeX command or,
2. Let novoalign add the M5 tags.

## 4.4 Paired End Alignment Mode

### 4.4.1 Scoring

Novoalign aligns paired reads against a reference genome using qualities and ambiguous nucleotide codes. The scoring system is based on Phred quality scores and the score for a paired alignment is  $-10\log_{10}(P(F | A_i))$  where  $P(F | A_i)$  is the probability that the fragment read by the sequencer originated from the alignment location.

A paired alignment score comprises three parts, Needleman-Wunsch alignment scores for each end of the pair in the form  $-10\log_{10}(P(R | A_i))$  and a fragment length penalty in the form  $-10\log_{10}(P(L | F))$



calculated from the fragment length distribution,  $F$ .

A posterior alignment score or quality is also given and is  $-10\log_{10}(1 - P(A_i | A_i, G, F))$  where  $P(A_i | A_i, G, F)$  is the probability of the alignment location given the read,  $R$ ; the genome,  $G$ ; and the fragment length distribution,  $F$ . For paired end reads the quality score is limited to not more than 150.

Setting of gap penalties and threshold is similar to single end novoalign.

## 4.5 Alignment process

With paired end reads Novoalign can have "proper fragments" and pairs that don't fit the fragment model.

The alignment process works as follows:

For Read1 Novoalign uses a seeded alignment process to find alignment locations each with a Read1 alignment score. For each good location found Novoalign does a Needleman-Wunsch alignment of the second read against a region starting from the Read1 alignment and extending 6 standard deviations beyond mean fragment length. The best alignment for Read2 will define the pair score for Read1/Read2. All the alignments are added to a collection for Read1.

This process is repeated using Read2 seeded alignment and then N-W for Read1, creating a collection of Read2/Read1 pairs. There are very likely duplicates amongst the two collections.

Novoalign then decides whether there is a "proper pair" or not. To do this a structural variation penalty is used as follows.

Novoalign has a proper pair if the score of the best pair (Read1/Read2 or Read2/Read1 combined score including fragment length penalty) is less than the structural variation penalty (default 70) plus best single-end Read1 score plus best single-end Read2 score.

If Novoalign has a proper pair, Read1/Read2 & Read2/Read1 lists are combined, removing duplicates and sorting by alignment score. At this point Novoalign has a list of one or more proper pair alignments. This list is passed to reporting which can report one or more alignments depending on the options.

If there wasn't a proper pair then Novoalign reports alignments to each read in single end mode and the reporting options will decide whether Novoalign reports one or more alignments.

The result of the paired search can be two paired alignments where the pairing is more probable than a structural variation, or it can be two individual alignments, one to each read of the pair.

Given the threshold, gap penalties and reads it is quite possible for novoalign to find alignments with gaps in both ends of the reads. There are no design restrictions that prevent this type of result and it depends only on the scoring parameters and threshold.

## 4.6 Bisulphite Mode

Bisulphite mode requires building of a double index, the first uses a hash table with all Cs translated to T's and the second a hash table with Gs translated to A's for fragments off the complementary strand.

Memory utilisation for the index may be higher in bisulphite mode than normal mode as we now have two hash tables. Novoindex will choose k & s values that allow the index to fit in RAM if possible. You can reduce memory further by increasing s or decreasing k.

Alignment is done iteratively gradually increasing error tolerance until a match is found. Each round of iteration will align the read in forward and reverse complement against the CT and the GA index. During CT alignment Cs in the read are translated to Ts for hash lookup, then during alignment, T's in the read can align to a T or a C in the reference sequence with no penalty. The process is then repeated for the GA alignment.

Scoring for alignments is similar to normal alignment scoring with difference that T in the read can align to a C in the reference without any penalty (or A to G for GA index alignments). This means that methylation status does not affect the alignment score.

In addition there is a command line option, -u, to impose a penalty on unconverted cytosines at CHG and CHH positions. If specified each unconverted cytosine in CHG or CHH positions in a read will be penalised thus biasing alignment in favour of methylated CGs.

The low-level of non-CpG methylation in vertebrates and the incomplete bisulphite conversion of unmethylated cytosines should be factored in to selecting this value. As a rough guide, a penalty can be worked out as follows:

Let  $P_{UC}$  be the probability an non-methylated cytosine is not converted,  $P_{CG}$  the probability that a cytosine at CpG is methylated and  $P_{CH}$  be the probability that a cytosine at a CHG or CHH is methylated. Then the probability of reading a cytosine at a CG position is:

$$P(C|CG) = P_{CG} + (1 - P_{CG}).P_{UC}$$

and the probability of reading a C at a CHN position is:

$$P(C|CH) = P_{CH} + (1 - P_{CH}).P_{UC}$$

We can then convert to log (phred) scale and calculate a penalty as:

$$\text{Penalty} = -10\log_{10}(P(C|CH)) + 10\log_{10}(P(C|CG))$$

Applying values from Ramsahoye et al. [6] for Drosophila

$$P_{CG} = 62\%, P_{CH} = 3\% \text{ (derived)}$$

and

$$P_{UC} = 1\%$$

$$\begin{aligned} \text{Penalty} &= -10\log_{10}(.03 + .97 * .01) + 10\log_{10}(.62 + .38 * .01) \\ &= -10\log_{10}(.04) + 10\log_{10}(.66) \\ &= 14 - 2 \end{aligned}$$

6 Ramsahoye BH, Biniszkiwicz D, Lyko F, Clark V, Bird AP, Jaenisch R. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl Acad. Sci. USA (2000) 97:5237–5242. [\[Abstract/Free Full Text\]](#)

As mentioned above, using a penalty for unconverted cytosines at CHG and CHH positions will slightly bias alignment in favour of methylated CG sites. This will mainly have an effect when there are multiple alignment sites with similar scores.

Novoalign will switch to bisulphite alignment mode whenever a bisulphite index is used.

### 4.6.1 Bisulphite Report Format

The differences to the output format are:

- 1) an indication of whether CT or GA index was used for the alignment. This is reported before mismatches and delimited from mismatches by a space.
- 2) Mismatches caused by unmethylated cytosines are shown with a hash '#' rather than a greater than '>' symbol. e.g. 5C#T to indicate a C in reference aligns to a T in the read and may be an unmethylated cytosine that was converted to uracil by bisulphite treatment. Similarly, 6G#A indicates a Cytosine on the complementary strand was unmethylated and hence appears in the read as an A.

The mismatch list does not show methylated cytosines as they match the reference sequence.

```
@chr2_98467308_98467344_1/1      S      CGGTATTGTAGAATAGTGTATATTAATGAGTTATAA
CBC??-@@BBBBBBB@@@BB??;??,6247092.  U      0      15      >chr2
98560453      R      .      .      .      GA 7G#A
@chr2_115989213_115989249_1/1      S      CGGTTTATTTTTTTTGGGGAATAGATTAAGTTTAAT
CCCCC-CCCCCCC== -?775BBB>>BC=B899;;9>  U      0      107     >chr2
116079348      F      .      .      .      CT 10C#T 13C#T
@chr2_48440862_48440898_1/1      S      CGGATATGTTATTTTAGGAGAAAAGAGGAAAAAATT
CCCCC=CCCCCCCCDD?BBC@CCCC@?2::<<022B  U      44      23      >chr2
48578844      R      .      .      .      GA 2T>A 9T>C 15G#A
@
```

## 4.7 Quality Calibration

Quality calibration is the process of re-evaluating base qualities using the actual counts of mismatches from alignments. The calibration in Novoalign is base specific which means two things:

1. We keep mismatch counts based on the actual base called so we can detect situations where, say, T is overcalled and likely to be wrong but calls of A, C & G are likely to be correct.
2. Rather than count “mismatches” we maintain counts for each of the bases aligned. This allows us to detect situation where a wrong call of, say, a T is more likely to be an A than a C. We can then calculate base specific mismatch penalties for each base at each position in a read.

These counts are used to calculate an actual mismatch probability or penalty as a function of: the position in the read; the “as called” base quality; the base called; and the base aligned. The empirical mismatch probability is then used in Novoalign alignment process in place of the “as called” base quality to set penalties for the alignment dynamic programming.

Categories used for counting mismatches are:

- The read within the pair (0 for first read, 1 for second read)
- The base position in the read, zero based.
- The “as called” quality
- The base called

For each combination, Novoalign maintains the count of the number of alignments to each of the four bases,  $M_A$ ,  $M_C$ ,  $M_G$  &  $M_T$ . Only ungapped alignments with a quality  $\geq 60$ , or  $\geq 70$  for paired end, are used to count mismatches.

The first step in the process of calculating calibrated qualities for each category involves binning counts across read length and quality values. Binning helps to increase the counts and to smooth fluctuations. Bins are 5 bases long and have variable number of quality values. At low qualities, bins take a single quality value, in mid range bins are 3 quality values wide and above a quality of 30 they are 5 wide. There is a bin for each base position and quality values so mismatch counts get added to multiple overlapping bins, this design eliminates edge effect between bins.

The second step involves adding priors to the count of calls and mismatches. Use of a prior helps stabilise calibrated quality values when counts are low. The prior is a minimum value for mismatch count and if the actual mismatch count is below the prior then we add extra mismatches to bring the count up to the prior and then a corresponding number of extra matches based on the “as called” quality. Unaligned reads (status NM) are also added to the priors as examples of correct base and quality calls.

Novoalign then calculates 4 base penalties is  $P_I = -10\log_{10}(M_I/N)$  for  $I$  in [ACGT] where  $M_I$  is the number of times an alignment matched base  $I$  and  $N$  is the total calls for this bin. The penalties are used in the dynamic programming alignment.

A Phred scaled quality value is also calculated as  $P = -10\log_{10}(M/N)$  where  $M$  is the total mismatches and  $N$  the total calls for the bin. This calibrated quality value is used in the report for the base qualities.

## 4.7.1 Using Quality Calibration

Quality calibration works for read files in the following formats:

- Solexa & Illumina FASTQ
- Sanger FASTQ
- FASTA Every base is assumed to have a starting quality of 30.
- FASTA with separate quality file
- BAM

The simplest way to use quality calibration is just to add the option **-k** to the Novoalign command line. This turns on calibration with calibration based on actual alignments. The calibration will start off neutral as a result of the priors and gradually, as more alignments are added, the calibration will shift to reflect the actual mismatch counts.

Novoalign also has the ability to save the mismatch count data and then use this as input to the calibration of a following run of Novoalign. Scenarios where this might be used include:

- Using mismatch counts from phiX lane to calibrate another lane
- Running an initial Novoalign at a low threshold to get mismatch statistics for use in a following run, possibly at a higher threshold. This would remove some startup effects from a single pass run.

Operation is controlled by two command line option:

- |                    |   |
|--------------------|---|
| <b>-k [infile]</b> | Enables quality calibration. The quality calibration data (mismatch counts) are either read from the named file or accumulated from actual alignments. Default is no calibration.<br>Note. Quality calibration does not work with reads in prb format.  |
| <b>-K [file]</b>   | Accumulates mismatch counts for quality calibration by position in the read and called base quality. Mismatch counts are written to the named file after all reads are processed. When used with -k option the mismatch counts include any counts read from the input quality calibration file. |

These two options can be used in several combinations :

- |                             |   |
|-----------------------------|---|
| <b>-k</b>                   | Turns on calibration with mismatch counting. Effects of calibration can be seen after a few thousand reads have been aligned. Calibration data is recalculated periodically as more reads are aligned.                                    |
| <b>-k <i>infile</i></b>     | Turns on calibration with mismatch counts read from <i>infile</i> . Mismatch counts from alignments are not used.   |
| <b>-K <i>outfile</i></b>    | Turns on mismatch counting without calibration. At the end of the run the mismatch counts are written to the <i>outfile</i> ready for use as input in another run.  |
| <b>-k -K <i>outfile</i></b> | Turns on calibration with mismatch counting. At the end of the run the mismatch counts are written to the <i>outfile</i> ready for use as input in another run. Calibration table is recalculated periodically as more reads are aligned. |

**-k *infile* -K *outfile*** Turns on calibration and mismatch counting. Initial mismatch counts are loaded from *infile*, new alignments are added to the counts, and then at the end of the run the mismatch counts are written to the *outfile* ready for use as input in another run. Calibration table is recalculated periodically as more reads are aligned.

### 4.7.2 Quality Calibration and Novoalign Reports

There is no change to the report format, for Novoalign the quality string displayed is now the calibrated qualities.

For Novoalign SAM format you can use the option `-rOQ` to add original quality tag `OQ:Z:qualities`

An R script 'qcalplot.R' that can produce charts of empirical quality for the reads from the mismatch file is included with the release.